Probabilistic cardiac and respiratory based classification of sleep and apneic events in subjects with sleep apnea

T Willemen[1,2,3], C Varon[2,3], A Caicedo Dorado[2,3], B Haex[1,4], J Vander Sloten[1] and S Van Huffel[2,3]

[1] Biomechanics section, Department of Mechanical Engineering, KU Leuven, Celestijnenlaan, Leuven, BE
[2] STADIUS, Department of Electrical Engineering (ESAT), KU Leuven, Kasteelpark Arenberg, Leuven, BE
[3] Medical Information Technologies, iMinds, Leuven, BE
[4] Maastricht University, Maastricht, NL

E-mail: tim.willemen@kuleuven.be

Abstract.

Current clinical standards to assess sleep and its disorders lack either accuracy or user-friendliness. They are therefore difficult to use in cost-effective population-wide screening or long-term objective follow-up after diagnosis. In order to fill this gap, the use of cardiac and respiratory information was evaluated for discrimination between different sleep stages, and for detection of apneic breathing. Alternative probabilistic visual representations were also presented, referred to as the hypnocorrogram and apneacorrogram. Analysis was performed on the UCD sleep apnea database, available on Physionet. The presence of apneic events proved to have a significant impact on the performance of a cardiac and respiratory based algorithm for sleep stage classification. WAKE versus SLEEP discrimination resulted in a kappa value of $\kappa = 0.439$, while REM versus NREM resulted in $\kappa = 0.298$ and light sleep (N1N2) versus deep sleep (N3) in $\kappa = 0.339$. The high proportion of hypopneic events led to poor detection of apneic breathing, resulting in a kappa value of $\kappa = 0.272$. While the probabilistic representations allow to put classifier output in perspective, further improvements would be necessary to make the classifier reliable for use on patients with sleep apnea.

PACS numbers: 87.85.Ng, 87.85.Tu

Keywords: biomedical signal processing, supervised learning, medical information systems, sleep research, Physionet

Submitted to: *Physiol. Meas.*

1. Introduction

Sleep is not an on-off state process, but is distributed over many different stages of sleep, alternating between episodes of Rapid-Eye-Movement (REM) sleep and NREM sleep, occasionally interrupted by episodes of WAKE (Rechtschaffen & Kales 1968).

Sleep apnea is a sleep-related breathing disorder which causes frequent arousal from sleep. It has an estimated prevalence of 14% in men and 5% in women, of which a large part still remains undiagnosed (Peppard et al. 2013). An apneic event is currently defined as a clear decrease from baseline in breathing volume of at least 10 seconds. Apneic events can be obstructive (complete or partial obstruction of upper airway with surrounding soft tissue) or central (absent or reduced breathing effort). In the case of only a partial obstruction, or when breathing effort is still present but significantly reduced, an apneic event is also referred to as a hypopnea (Quan et al. 1999).

Currently, there are two main standards to assess sleep and its pathologies. One is the current gold standard, polysomnography (PSG), but lacks ease of use towards both patient and physician due to the required attachment of numerous sensors. Annotations of sleep stages and/or apneic events are performed manually, resulting in inter- and intra-observer variability (Whitney et al. 1998). The second standard, actigraphy, is a very unobtrusive and user-friendly method, but lacks accuracy due to its simple analysis based on movement activity (Morgenthaler et al. 2007) alone. Research has however shown a significant influence of sleep and its disorders on the autonomic nervous system, of which sympathetic and parasympathetic variations can be analyzed using cardiac and respiratory changes (Somers et al. 1993, Burgess et al. 1997). The current gap in clinical standards towards assessment of sleep and its pathologies could thus be filled using automated cardiac, respiratory and movement based analysis of sleep.

Our previous work has already shown promising results in distinguishing between different stages of sleep in healthy adults (Willemen et al. 2014a). Furthermore, on a dataset containing a wide range of adults with different levels of apnea severity, we reached an accuracy above 90% in distinguishing between apneic and healthy breathing episodes (Willemen et al. 2014b). Apneic events have however a significant impact on both the respiratory and cardiac system. This study will therefore evaluate this influence of apneic events on the ability to use cardiac and respiratory information for discriminating between WAKE, REM, light (N1N2) and deep (N3) NREM sleep. Discrimination between APNEIC and HEALTHY breathing will also be assessed. Moreover, an alternative visual representation of the classifier output is presented, exploiting the benefit of a machine learning approach compared to the gold standard discrete manual annotation of both sleep stages and apneic events.

2. Methodology

2.1. Database

The database used in this study (available online at Physionet) is entitled St. Vincent's University Hospital and University College Dublin Sleep Apnea Database (UCD database) (St. Vincent's University Hospital/University College Dublin 2008), and contains 25 overnight PSG recordings of subjects suspected of sleep-disordered breathing (21 males and 4 females). An overview of the general properties on the subjects and recordings within the database can be found in Table 1. Expert reference annotations were given on the presence of apneic events according to clinical standards (onset and duration of every apneic/hypopneic event), as well as on sleep stages for every 30-second epoch (window) according to Rechtschaffen and Kales (R&K) rules: WAKE, REM, NREM (Stage 1, Stage 2, Stage 3, Stage 4).

Since the more recent American Academy of Sleep Medicine (AASM) standard for

Table 1. General properties on subjects and recordings within the UCD database.

	Mean ± Std	**Range**
Age	49.96 ± 9.55 years	28.0 - 68.0 years
BMI	31.60 ± 4.03 kg/m^2	25.1 - 42.5 kg/m^2
AHI	24.24 ± 20.29	2.0 - 91.0
Duration of measurement	6.93 ± 0.54 hours	5.9 - 7.7 hours
Sleep efficiency	77.20 ± 11.16 %	58.0 - 92.0 %

Table 2. Overview of the amount of 30-second epochs within each category (absolute amount and relative percentage). Values given are the mean, standard deviation and range over the 25 recordings.

	Mean amount of epochs		**Range of the amount**	
	#	%	#	%
APNEIC	178.6 ± 134.1	21.9 ± 17.7	15 - 593	1.8 - 83.4
HEALTHY	653.1 ± 159.1	78.1 ± 17.7	118 - 856	16.6 - 98.2
WAKE	189.0 ± 85.4	22.9 ± 11.0	71 - 357	8.2 - 41.6
REM	120.6 ± 63.4	14.5 ± 7.6	19 - 246	2.3 - 29.4
N1N2	415.5 ± 87.1	50.0 ± 9.8	276 - 592	34.2 - 69.3
N3	106.5 ± 57.3	12.7 ± 6.7	0 - 248	0.0 - 30.7

NREM sleep stage annotations makes use of annotations N1 through N3, the original R&K Stage 1 through Stage 4 annotations were mapped to the new N1 through N3 annotations according to AASM guidelines, i.e. Stage 3 and Stage 4 were combined in N3, while Stage 1 and Stage 2 correspond respectively to N1 and N2. Next, N1 and N2 annotations were combined to represent light sleep, whereas N3 represents deep sleep. The event based apneic annotations were transformed to 30-second epoch based annotations according to the following rule: if at least one second of the 30-second epoch has an overlap with an event based apneic annotation, the epoch is considered as APNEIC, whereas in the case of no overlap it is considered as HEALTHY. An overlap of only one second was chosen to make up for the fact that the event based annotations do not cover the sequential recovery breaths and autonomic influences, but only the periods of reduced breathing volume. An overview of the amount of epochs within each category can be found in Table 2. The total amount of epochs in the complete data set equals 20793.

2.2. Data preprocessing

Cardiac activity was extracted from the Electrocardiogram (ECG), while respiratory activity was extracted from the ribcage breathing belt signal. The PSG recordings did not contain any signal directly related to movement information. Interbeat RR intervals were extracted from the 128 Hz ECG using the Pan-Tompkins algorithm (Pan & Tompkins 1985). To identify and correct false positive and false negative R-peak detections, a search-back post-processing algorithm was applied, as proposed in (de Chazal et al. 2004). The resulting interbeat RR interval series will from now on be labelled as the RR time series. The 8 Hz ribcage breathing belt signal was stripped from baseline wander by subtracting a 20 second median filtered version of the original signal (± 5 average breathing intervals). The resulting signal will from now on be

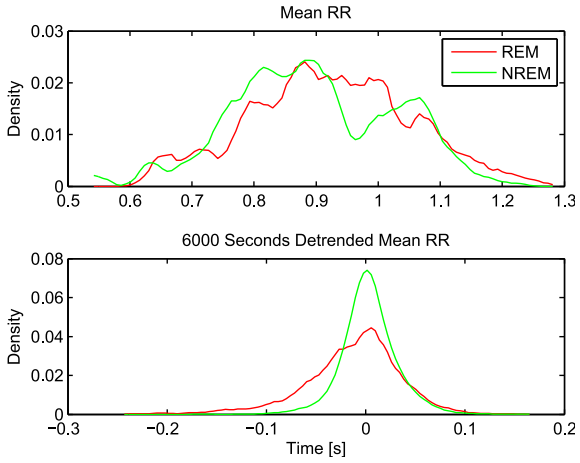

Figure 1. Example effect of detrending on the probability density function

labelled as the BS signal. Individual breathing intervals were extracted by locating subsequent peak and valley positions. From these starting locations of breathing in (valleys) and breathing out (peaks) intervals, two time series were extracted: Firstly the interbreath interval series (from now on labelled as BR time series), and secondly the inspiration-to-expiration ratio interval series (from now on labelled as IEratio time series).

In order to account for inter-subject variability, three detrended versions of the RR, BR and IEratio were defined for the overnight data of every subject. Firstly, 6000 seconds (100 minutes) corresponding on average to one NREM-REM sleep cycle. Secondly, 600 seconds (10 minutes) reflecting shorter changes between light and deep sleep, or light and REM sleep. And lastly, 120 seconds (2 minutes) reflecting short awakenings or arousals. Detrending was performed by subtracting a median filtered version of the respective window size from the different time series. An example of the effect of detrending on the probability density function is displayed in Fig. 1. Detrending over 6000 seconds enables to filter out the inter-subject variability, as well as the daily circadian variation in average heart rhythm.

The resulting time series from the overnight data of every subject were segmented in epochs of 60 seconds with 30 seconds overlap. An overlap of 30 seconds was chosen to enable the future classifier annotations on the 60-second epochs to be compared with the reference annotation on the 30-second epochs, which will be discussed in more detail later on. An epoch length of 60 seconds was chosen because it achieved the best results in distinguishing apneic from healthy breathing in a study comparing different epoch lengths, ranging from 15 seconds to 90 seconds (de Chazal et al. 2004). Moreover, in our previous work we also opted for an epoch length of 60 seconds instead of using the gold standard size of 30 seconds as used in the reference Electroencephalography (EEG) based sleep stage annotations (Willemen et al. 2014a). The predominant frequencies in our EEG brain waves are an order of two greater than the frequencies observed in our heart and breathing rhythm, making the choice for 30-second epochs in cardiac and respiratory based sleep stage scoring less obvious. The Heart Rate Variability (HRV) Task Force also recommends interval lengths of at least 10 times the wavelength of the lowest frequency bound (Malik et al. 1996), making

Table 3. Overview of time and frequency domain features, numbered between brackets

Time domain	(1) Mean; (2) Percentiles (p = 10,25,50,75,90); (3) Inter quartile/decile range (IQR); (4) Standard deviation (STD); (5) Standard deviation of interbeat differentials (STDi); (6) Mean absolute deviation (MAD1); (7) Median absolute deviation (MAD2); (8) Root mean square of interbeat differentials (RMSi); (9) Serial correlation coefficient (k = 1,2,3,4,5) (SerCorrCoef)
Frequency domain	(10) VLF, LF, HF and TF energy content (EC); (11) LF to HF energy content ratio (ECR); (12) VLF, LF and HF normalized energy content (EC norm); (13) Mean frequency; (14) Peak frequency; (15) Cumulative energy frequency percentile (CE freq perc) (p = 25,50,75); (16) Cumulative energy frequency inter quartile range (CE freq IQR)

60 second intervals at least reliable for the HF interval (0.15 Hz - 0.40 Hz). For the LF and VLF interval, 60 second intervals are considered too short to be fully reliable, which is an important limitation of this approach. Another limitation with respect to accurate HRV analysis is the low sample rate of the ECG signal at 128 Hz, which will have introduced some quantification noise to the RR time series. This noise can be reduced by interpolating the ECG signal to 1000 Hz before actual R-peak detection.

2.3. Feature extraction

For every 60-second epoch, a multitude of time and frequency domain features were calculated, of which an overview can be found in Table 3. Exact feature definitions can be found in other literature (e.g. in (Bsoul et al. 2011)). All of the listed time domain feature types (1-9) were calculated for the original RR time series (18 features). The first three feature types were also calculated on the three detrended RR time series (3 x 8 = 24 features). For the original BR and IEratio time series, feature type 9 was skipped due to the limited amount of points within each window (13 features each). The first 3 feature types were again also calculated on the three BR and IEratio detrended time series (3 x 8 = 24 features each). For the BS signal, feature types 5, 8 and 9 were skipped since they would mostly reflect technological characteristics instead of physiological characteristics (11 features). Frequency domain features (10-16) were calculated for every 60-second epoch of the 120 seconds detrended 4 Hz interpolated RR and BR time series, and for every 60-second epoch of the BS signal (14 features each). Before every Fast Fourier Transform (FFT) calculation, a Hanning window was applied to suppress spectral leakage. Frequency domain features were deemed irrelevant on the IEratio time series, and therefore skipped. Overall, this leads to a total amount of 184 features: 56 RR features, 51 BR features, 37 IEratio features and 25 BS signal features.

Next, to improve Gaussianity of features, Box Cox transformed features were defined for every feature type that can only be positive (thus for all feature types except 1, 2 and 9), defined as

$$\boldsymbol{d}_{BoxCox} = \begin{cases} \frac{\boldsymbol{d}^{\lambda}-1}{\boldsymbol{\lambda}}, & \lambda \neq 0 \\ \ln(\boldsymbol{d}), & \lambda = 0 \end{cases} \quad (1)$$

with a feature-specific optimized λ that maximizes the Log-Likelihood Function given the feature input values $\boldsymbol{d}$. To account for inter-subject variability, night-specific normalized features were also defined for every feature (including the Box Cox transformed features),

$$\boldsymbol{d}_{norm} = \frac{\boldsymbol{d} - \text{median}\,(\boldsymbol{d})}{\text{mad}\,(\boldsymbol{d})}, \tag{2}$$

with median and median absolute deviation (mad) specific for every feature and overnight measurement. These two transformations increase the total amount of features significantly, leading to a total feature set of 510 features.

2.4. Feature classification

A total of five binary classification problems were assessed, namely APNEIC versus HEALTHY breathing, WAKE versus SLEEP at sleep onset (first hour of the night), WAKE versus SLEEP after sleep onset (all but first hour of the night), REM versus NREM and N1N2 versus N3. The results of the latter four were also combined to evaluate the overall classification problem of WAKE-REM-N1N2-N3. The problem of WAKE versus SLEEP was split up in two parts, since the physiological changes occurring around sleep onset are different from awakenings during the night.

Since the reference annotations are linked to 30-second epochs, every 60-second epoch (on which features were calculated) essentially contains two reference annotations. To determine the annotation of the 60-second epoch (to be used in the classifier training phase), both 30-second epoch reference annotations were combined, prioritizing in the case of a tie APNEIC over HEALTHY breathing, WAKE over SLEEP, REM over NREM and N1N2 over N3. For example, if a 60-second epoch contains the annotations N2 and REM, it will be referred to as REM.

Classification was performed using Linear Discriminant Analysis (LDA) (Ripley 1996). For a given (to be classified) input feature vector $\boldsymbol{x}_i$, representing the features defined on a 60-second epoch, the trained LDA classifier (with class-specific prior probability parameters π_k and covariance matrix regularization parameter α) will output two discriminant values y_1 and y_2 (one for each class k). From these two discriminant values, posterior class probability estimates can be derived using

$$P\,(k\,|\boldsymbol{x}_i\,) = \frac{\exp\,(y_k)}{\sum_{l=1}^{2}\exp\,(y_l)}. \tag{3}$$

The class of the input feature vector $\boldsymbol{x}_i$ is then determined as the class with the highest posterior probability estimate.

Training vectors $\boldsymbol{x}_{nk}$ were first ridded from outliers within their feature values. Outlier values were replaced by their class-specific outlier bounds, defined as respectively the 25^{th} and 75^{th} percentile $\pm$ 1.5 times the inter quartile range (IQR). Next, training vectors were normalized to the range [-1,1]. The same normalization to the range [-1,1] was also applied on every input feature vector $\boldsymbol{x}_i$ before classification.

Due to the 30 second overlap of the 60-second epochs, every 30-second epoch linked to the reference annotations is essentially scored twice by the 60-second epoch classifier. In order to take both classifications into account, the posterior class probability estimates from both of the 60-second epochs were averaged, resulting in 30-second epoch posterior class probability estimates. To filter out rapid transitions

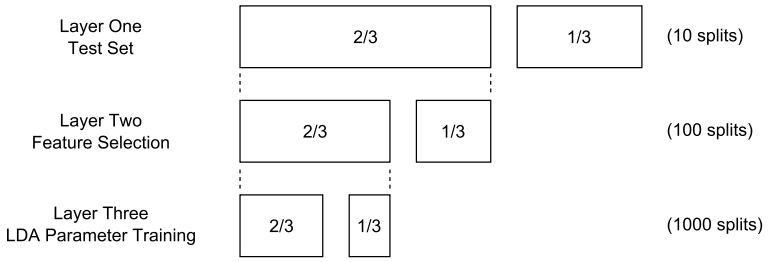

Figure 2. Visualization of the triple layer validation scheme

between APNEIC and HEALTHY breathing, or between different sleep stages, the 30-second epoch posterior class probability estimates were averaged over those from their two neighbouring epochs. The final class of the 30-second epoch classifier annotation was then determined by the highest average probability estimate.

To prevent overfitting of the data, a triple layer validation scheme was used, as visualized in Fig. 2. In the first layer, the 25 nights of the dataset were split up via stratified sampling into on average 2/3 training nights and 1/3 test nights. Stratified sampling makes sure that a certain aspect within the data is equally represented. The goal of this study is to assess the influence of apneic events on the ability to use cardiac and respiratory information for sleep staging. Therefore, a more or less equal ratio of APNEIC to HEALTHY epochs in each training and test set was taken as the stratified sampling constraint. This process was repeated ten times, in order to provide robust measures of classifier performance, which are less affected by choice of training and test nights. For all of the five binary classification problems, the same ten distributions of training and test nights were used.

In the second layer, each set of training nights from the first layer was similarly split up ten times via stratified sampling into on average 2/3 training nights and 1/3 validation nights, resulting in a total of 100 splits. This second layer allowed for a robust selection of ten feature subsets, one for each of the ten folds of the first layer. A feature subset was constructed based on a maximum value of kappa (Landis & Koch 1977) on the validation nights, averaged over the ten folds of the second layer. Feature selection will be further discussed in section **??**.

A similar third layer (resulting in a total of 1000 splits) allowed for a robust gridsearch optimization of the LDA classifier parameters (α, π_1, $\pi_2 \in [0,1]$) to be used in the respective fold of the second layer. Optimal parameters were selected based on a maximum value of kappa (Landis & Koch 1977) on the validation nights, averaged over the ten folds of the third layer. The values of the classifier parameters to be used in the respective fold of the first layer were derived from the selected parameters used within the second layer, more specifically by taking the median of their values.

Classifier performance was also assessed using a leave-one-out validation scheme (splitting up the dataset 25 times in 24 training nights, and 1 test night), resulting in a night-by-night evaluation of classifier performance. Posterior class probability estimates were averaged over the outputs of the ten constructed classifiers from the triple layer validation scheme (each with their own optimized feature subsets and classifier parameters).

2.5. Feature selection

In order to reduce complexity and noise on the classification result, unique subsets of features were selected as being most descriptive for their respective binary classification problems. To reduce computational cost, the feature selection process was performed in two phases. In phase one, classifier performance was assessed (on layer two) for each of the 510 features separately. From the feature sets of each of the four time series or signals (RR, BR, IEratio and BS), only the top 25 features achieving the highest kappa value were kept, reducing the feature set to 100 features. In phase two, a sequential forward floating search (SFFS) algorithm (Pudil et al. 1994) extracted a subset of maximum ten features from this reduced set based on a maximal increase in kappa (on layer two) after every step. The SFFS algorithm alternates between forward selection and unselection of one of the previously selected features, as long it leads to an increase in kappa. If at a certain point neither the forward selection nor the unselection of a previously selected feature leads to an improvement in kappa, the algorithm stops and outputs the final feature subset. The algorithm also stops if the maximum of ten features is reached.

During each forward selection step of the SFFS algorithm, in order to prevent selection of too similar features (and to prevent problems due to multicollinearity), selectable features from the remaining feature set were assessed based on an inter-feature mutual information (MI) criterion. The inter-feature mutual information measures the information shared between two features. The criterion states: 'Whenever the inter-feature mutual information of a selectable feature with any feature from the already selected feature set was higher than an heuristically determined threshold of 1.246, the selectable feature was disregarded from the remaining feature set.' The threshold of 1.246 was determined by calculating the inter-feature mutual information for all possible combinations of the 510 features on the complete dataset (a total of 510 x 510 = 260100 values). The resulting values were sorted and the elbow criterion applied. The elbow point coincided with the 85 percentile, of which the corresponding mutual information value of 1.246 was thus taken. This means that every feature will have on average 15% features that are deemed too similar.

2.6. Visual representation

The gold standard technique of sleep stage and apneic event annotation is done manually, leading to a discrete set of decisions on the current sleep stage, and the presence or absence of apneic breathing. Several limitations to this gold standard approach are discussed by Himanen & Hassan (2000). It is well known that a significant disagreement ratio exists when PSG's are annotated by different sleep experts (Whitney et al. 1998). This can be due to ambiguous data, ambiguous interpretation of the AASM standard, or just due to simple human error. Opportunities of computer based sleep recording and analysis are extensively discussed by Penzel & Conradt (2000a) and Schulz (2008). Advantages of a machine learning approach are not only the consistency in decisions, but also the possibility to extract class probability estimates. A machine learning approach can thus e.g. give an indication of how certain it is about its decision, based on the information it has gathered on other subjects and nights during its training phase.

An alternative, non-discrete visual representation of the classifier output is therefore proposed, referred to as the hypnocorrogram (for the distribution of sleep

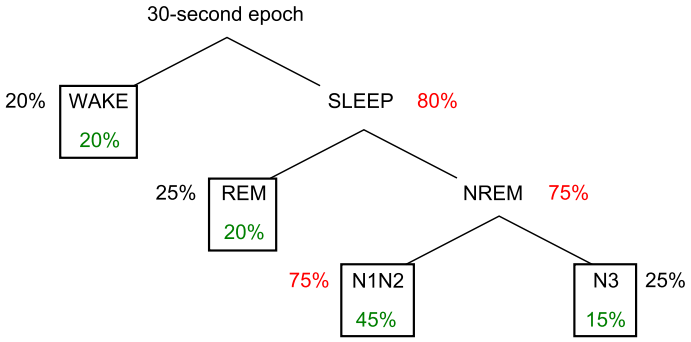

Figure 3. Combining binary classifier posterior probability estimates to attain multiclass posterior probability estimates

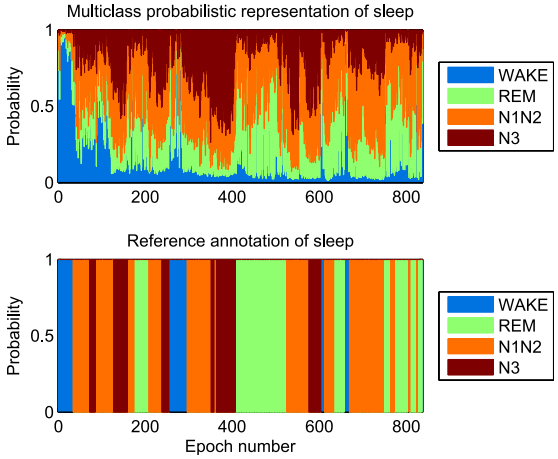

Figure 4. An example hypnocorrogram, visualizing the multiclass probabilistic representation of the different sleep stages.

stages during the night) or the apneacorrogram (for the presence of apneic breathing). By extracting the posterior class probability estimates for each of the four sleep stage classification problems, the outputs can be combined using a probabilistic decision tree, as visualized in Fig. 3. This thus leads to four multiclass posterior probability estimates, which can be plotted simultaneously on a nightly basis as in Fig. 4. The different coloured areas represent the magnitude of the probability at every 30-second epoch during the night. The reference annotation (with only 0% or 100% posterior probability values) is displayed below for comparison.

Fig. 5 gives a probabilistic representation on the presence or absence of apneic breathing, referred to as the apneacorrogram. The coloured area above the zero-line represents the magnitude of the probability above 50% (i.e. the classifier leaning more towards APNEIC breathing), while the coloured area below the zero-line represents the magnitude below 50% (i.e. the classifier leaning more towards HEALTHY breathing). The reference annotation is displayed below for comparison. The content of the figures will be further discussed in section 3.3.

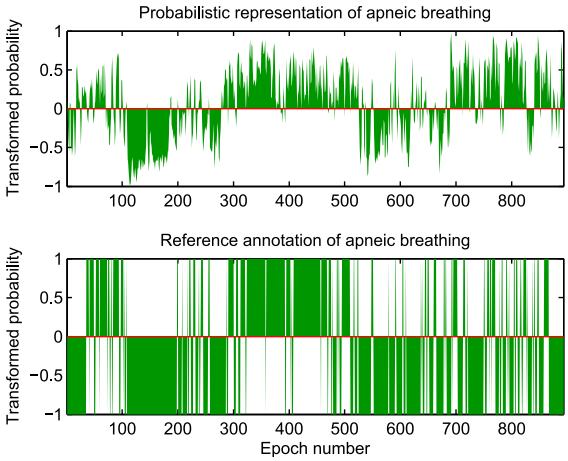

Figure 5. An example apneacorrogram, visualizing the probabilistic representation of apneic breathing

3. Results

3.1. Feature selection

Due to the large feature starting set of 510 features, a lot of different features were selected at least once. Combined over all classification problems, most features were derived from the BS signal (33%), followed by the RR series (32%), IEratio series (20%) and lastly the BR series (15%). More than half of all selected features (56%) were normalized versions of features, proving the added value of this transformation. On the features where a Box Cox transform could have been defined, it was present in 44%. Since LDA makes the assumption that the probability density functions of features are normally distributed, applying the Box Cox transform should always be considered.

In order to determine the relevance of selected features present in one or more of the ten different feature subsets extracted by the triple layer validation scheme, selected features were given a feature score (separately for every binary classification problem). Whenever a feature was present in one of the feature subsets, its feature score was increased by a number between 1 and 10 based on how early it was selected by the SFFS algorithm (10 for the first feature, 9 for the second feature, etc.). Since ten different feature subsets exist, the maximum feature score attainable was 100. Table 4 lists the few selected features with a feature score above or equal to 20.

For the APNEIC vs HEALTHY breathing classification problem, features related to the variability in the BR time series were selected most frequently, indicating that during apneic events the breathing rate becomes slower than normal, while during the recovery breaths the breathing rate becomes faster. For the WAKE vs SLEEP classification problem, the most selected features were related to the high and low frequency energy content in the BS signal, indicating a more controlled and constant breathing rhythm and volume during sleep compared to wake, free from regular movement artifacts (which affect the breathing belt signal). For the REM vs NREM classification problem, the most selected features were either related to a higher heart rate and breathing rate rhythm during REM sleep (as calculated on

Table 4. Overview of selected features with a feature score of at least 20. EC norm = energy content normalized; ECR = energy content LF/HF ratio.

Feature score	Feature description	Mean ± std	
	APNEIC vs HEALTHY	**APNEIC**	**HEALTHY**
78	Box Cox VLF EC of BR	-1.77 ± 1.06	-2.86 ± 1.14
30	Box Cox MAD2 of BR	-0.33 ± 0.41	-0.79 ± 0.49
20	Box Cox MAD1 of BR	-0.62 ± 0.45	-1.08 ± 0.43
	WAKE vs SLEEP (first hour)	**WAKE**	**SLEEP**
44	Normalized LF EC norm of BS	8.20 ± 10.13	-0.35 ± 0.88
31	LF EC norm of RR	0.34 ± 0.17	0.42 ± 0.22
20	Normalized Box Cox ECR of BS	1.67 ± 1.97	-0.36 ± 1.07
	WAKE vs SLEEP (after first hour)	**WAKE**	**SLEEP**
70	HF EC norm of BS	0.86 ± 0.12	0.96 ± 0.02
32	Normalized ECR of BS	6.42 ± 9.23	-0.24 ± 0.92
	REM vs NREM	**REM**	**NREM**
57	Normalized 25th perc of 6000s detrended RR	-0.35 ± 1.84	0.35 ± 0.98
40	Normalized VLF EC norm of RR	0.69 ± 1.46	-0.24 ± 1.06
27	Normalized Box Cox HF EC norm of RR	-0.61 ± 1.35	0.34 ± 1.13
20	Normalized mean of 6000s detrended BR	-0.61 ± 1.92	0.15 ± 1.04
	N1N2 vs N3	**N3**	**N1N2**
22	Normalized Box Cox VLF EC of BR	-0.98 ± 0.93	0.13 ± 1.19
20	Box Cox STD of BR	-0.98 ± 0.27	0.46 ± 0.52
20	Box Cox ECR of BS	-4.28 ± 0.54	-3.41 ± 1.00

the 6000 seconds detrended RR and BR time series which average out the circadian variation), or related to a higher amount of low frequency and a lower amount of high frequency content in the RR time series during REM sleep (indicating a higher amount of sympathetic activity). For the N3 vs N1N2 classification problem, the most selected features were related to the variability in the BR times series and BS signal, indicating a more controlled and constant breathing rhythm and volume during deep sleep (N3) compared to light sleep (N1N2). However, the large amount of apneic events during light sleep compared to deep sleep might have also influenced the selection of these features.

3.2. Classifier performance

The classifier performance for each of the classification problems is listed in Table 5 for both the triple layer validation scheme as the leave-one-out validation scheme. Performance metrics given are the mean, standard deviation and range (over the different folds) of kappa, agreement, precision and sensitivity. Agreement values are around 75% to 80% for all binary classification problems, indicating reasonable performance at first sight. Sensitivity is however clearly lacking, especially for the REM vs NREM classification problem (<40%). Precision values are only around 40% to 45% for all binary classification problems. The low sensitivity and precision are clearly reflected in the low values of kappa, with values between 0.20 and 0.40 only considered as fair agreement, and values above 0.40 as moderate agreement (Landis & Koch 1977). The range of values in the triple layer as well as the leave-one-out scheme indicate the considerable effect the choice of training and test nights can have on measures of classifier performance.

Table 5. Classifier performance using the triple layer and leave-one-out validation scheme. Kappa has been scaled with a factor 100. Other results are given in percentages. The range of values is displayed between brackets.

	Kappa	**Agreement**	**Precision**	**Sensitivity**
Triple layer				
APNEIC vs HEALTHY	30.24 ± 5.15 [24.70 - 40.46]	73.58 ± 5.20 [63.47 - 79.52]	42.93 ± 4.50 [36.48 - 48.33]	59.52 ± 15.19 [34.66 - 75.87]
WAKE vs SLEEP	37.07 ± 11.21 [13.66 - 47.31]	74.22 ± 6.23 [60.89 - 79.30]	46.07 ± 8.38 [31.04 - 55.98]	71.66 ± 6.97 [62.18 - 88.00]
REM vs NREM	23.50 ± 5.93 [16.46 - 34.58]	77.28 ± 3.04 [70.86 - 81.75]	38.37 ± 7.54 [27.95 - 51.02]	38.52 ± 9.33 [27.34 - 55.59]
N3 vs N1N2	29.86 ± 5.99 [18.89 - 39.78]	73.27 ± 2.94 [68.00 - 77.29]	39.14 ± 6.45 [26.69 - 52.24]	62.92 ± 5.96 [53.96 - 75.15]
Leave-one-out				
APNEIC vs HEALTHY	27.21 ± 17.92 [-11.48 - 55.90]	75.90 ± 8.75 [48.97 - 90.63]	41.13 ± 22.75 [2.13 - 88.03]	62.30 ± 21.18 [3.70 - 100.00]
WAKE vs SLEEP	43.91 ± 19.30 [-7.55 - 74.93]	77.95 ± 9.36 [46.41 - 91.04]	51.98 ± 20.20 [24.01 - 88.89]	79.46 ± 10.62 [54.55 - 95.65]
REM vs NREM	29.80 ± 22.99 [-8.69 - 70.66]	81.52 ± 7.46 [62.01 - 93.48]	45.62 ± 23.20 [0.00 - 85.53]	39.38 ± 20.45 [0.00 - 78.65]
N3 vs N1N2	33.90 ± 23.15 [-27.10 - 66.59]	76.92 ± 12.60 [38.34 - 96.22]	44.01 ± 19.67 [0.00 - 77.87]	69.99 ± 20.84 [1.92 - 100.00]

Table 6. Confusion matrix of the combined sleep stage classification problem, more specifically the sum of the confusion matrices from each of the 25 nights/folds of the leave-one-out scheme. The columns represent the predicted annotations by the classifier (indicated as P). The rows represent the actual reference annotations (indicated as A).

A ↓ \ P →	**Wake**	**REM**	**N1N2**	**N3**
Wake	2870	276	1239	341
REM	381	956	1401	278
N1N2	1372	710	6173	2133
N3	83	41	694	1845

To assess the combined sleep stage classification problem, Table 6 shows the complete confusion matrix, built up by adding the confusion matrices from the individual folds of the leave-one-out validation scheme. The total sum of elements adds up to the total amount of epochs present in the dataset (=20793). As seen from the confusion matrix, the classifier has a lot of difficulties in distinguishing Wake, REM and N3 from the most common state N1N2, leading to mediocre overall performance values.

3.3. Visual representation

Fig. 4 shows the hypnocorrogram of the combined sleep stage classification problem for the night of a subject with an AHI index of 14. In order to visualize the proposed representation in a sufficiently clear way, it is an example with a kappa performance value of the WAKE-REM-N1N2-N3 problem on the upper end of the spectrum (44.28 to be exact). Even then, it is quite clear that the classifier is in most cases far from certain about its decisions, which is exactly the point of this representation. These classifiers were built to assess how unseen data corresponds to training data. They should therefore exactly output how likely this unseen data corresponds to each of the classes within the training data, without pretending to know which class it would fit in the end. The same is true for Fig. 5, representing the apneacorrogram of APNEIC vs HEALTHY breathing. The information present in the uncertainty of the classifier is valuable and worth visualizing.

4. Discussion

This study evaluated the use of cardiac and respiratory information for discriminating between the different sleep stages WAKE, REM, light (N1N2) and deep (N3) NREM sleep, as well as discriminating between APNEIC and HEALTHY breathing. Moreover, it presented an alternative visual representation of the classifier output, the hypnocorrogram and apneacorrogram, incorporating the uncertainty of the classifier. The authors believe that visualizing the correspondence of unseen data to training data allows to put the classifier output into perspective, and will increase the likelihood of bringing these kinds of classifiers into clinical practice.

Table 7 compares the obtained classification results to similar classification algorithms in literature. The last row of the table lists the presented results of this study. The top two rows present our previous results on classification of sleep stages in healthy subjects (Willemen et al. 2014a), and on classification of apneic breathing (Willemen et al. 2014b). During our literature research, the study of Redmond & Heneghan (2006) was the only one found on cardiac and respiratory based classification of sleep stages in apnea patients, and was therefore added to Table 7. Their used database was however not publicly available. In total, five studies were found using the same public UCD database as in this study. Two of them however used the UCD database for evaluation of pulse oximetry based or EEG based classification, which was not part of this work. Two others studied the presence of apneic breathing using cardiac and respiratory information, but added the UCD database to a larger non-public database, making direct comparison impossible. The fifth study, from Xie & Minn (2012), did use only the UCD database for their cardiac and respiratory based classification of apneic breathing, and was therefore also included in Table 7.

As shown in Table 7, our results on the apnea classification problem are significantly worse in this study, compared to our previous study (Willemen et al. 2014b) which used the Physionet Apnea-ECG database (Penzel et al. 2000b). Although there were some differences in feature extraction, selection and classification methodology between both studies (e.g. epoch length), these could not fully explain the large differences achieved between the two databases. The feature set and classifier algorithm of Bsoul et al. (2011), which reached an agreement of 86% on the Apnea-ECG database, also only achieved an agreement of 71% on the UCD database. The latter result was reported by Xie & Minn (2012), who used exactly the same feature

Table 7. Overview of similar classification algorithms in literature. Target: Sleep = WAKE-REM-NREM; Apnea = APNEIC vs HEALTHY. Signals: EDR = ECG derived respiration; RIP = respiratory inductance plethysmography (breathing belts); MOV = movement. Nights: h = healthy; a = apneic.

Author	Signals	Target	Agreement	Kappa	Nights	Epoch
Willemen et al. (2014a)	RR/RIP/MOV	Sleep	81%[a]	0.62[a]	85 h	60s
Willemen et al. (2014b)	RR/EDR	Apnea	90%[a]	0.80[a]	70 a	60s
Redmond & Heneghan (2006)	RR/EDR/RIP	Sleep	67%[b]	0.32[b]	37 a	30s
Xie & Minn (2012)	RR/EDR	Apnea	78%[c]	-	25 a	60s
This study	RR/RIP	Apnea Sleep	74%[d]/76%[b] 65%[d]/70%[b]	0.30[d]/0.27[b] 0.34[d]/0.41[b]	25 a	30s

[a]Result on separate test set of 35 nights
[b]Averaged result of leave-one-out cross-validation
[c]Averaged result of ten-fold cross-validation (possible night-specific training)
[d]Averaged result of triple layer validation scheme

set of Bsoul et al. (2011), but evaluated it on ten different classifier algorithms. On a side note, reported agreements by Xie & Minn (2012) ranged between 68% and 78%, showing a significant influence of the classifier algorithm on the overall agreement. Although it is known that supervised learning algorithms can have significant variability in their performance across problems (Caruana & Niculescu-Mizil 2006), the large differences in agreement found by Xie & Minn (2012) could also be explained by their reported use of default parameter settings for all but one classifier.

The difference between the Apnea-ECG and UCD database might be explained by differences in apnea severity of the subjects contained within the databases. The Apnea-ECG database contains a total of 13002 apneic events and 1337 hypopneic events. The UCD database however contains only 695 apneic events, but a total of 2623 hypopneic events. This means that the ratio of hypopneic to apneic events in the UCD database is 36.7 times larger than the same ratio in the Apnea-ECG database. Since hypopneic events are the less-severe variant of apneic events, they have less impact on the cardiac and respiratory system, and are thus more difficult to detect. This is proven by the fact that a total of 92.8% of all apneic events were detected in the UCD database using the leave-one-out cross-validation scheme, while only 61.4% of hypopneic events were detected.

Another large difference between the two databases are the event based apneic annotations. While the latter normally only contain periods of reduced breathing volume (as in the UCD database), the annotations in the Apnea-ECG database also contain the associated recovery breaths (de Chazal et al. 2004). The use of 60-second epochs with 30-second overlap made sure that features from the recovery breaths were also linked to the apneic reference annotation epochs in the UCD database. However, this information will also have been linked to healthy reference annotation epochs, obscuring the decision boundary between apneic and healthy. A solution might thus be to artificially increase the length of each of the event based apneic annotations of the UCD database (for example by 10 seconds) as to include at least the first recovery

breaths into the annotation.

The results of our sleep stage classification problem are similar to those reported in the study by Redmond & Heneghan (2006). They are however significantly worse when compared with our previous study on sleep stage classification in healthy subjects (Willemen et al. 2014a). This can be due to several reasons. Firstly, our previous study also included features related to movement, as measured by indentation sensors inside of the bed system (Verhaert et al. 2011). These movement features proved to be very selective, not only in classification of wake epochs, but also in distinguishing light sleep (N1N2) from deep sleep (N3) (long-term absence of movement activity during deep sleep), as well as REM sleep from NREM sleep (short twitches present in REM sleep). A signal related to movement was however not present in the UCD database. Secondly, the presence of apneic events in both REM as NREM sleep will have definitely obscured the decision boundary between both, since apneic events have a similar impact on the cardiac and respiratory system independently of the sleep stage. An improved accuracy might be obtained if epochs which overlap with apneic events would be ignored during the training phase of the classifier. The general higher sympathetic activity perceived in apnea patients (Somers et al. 1995) will have also influenced the limited bias in classifier output. Thirdly, The frequent arousals from sleep due to apneic events obscured the decision boundary between wake and sleep, since the effect of an arousal on the heart rhythm is quite similar to the effect of an awakening.

Redmond & Heneghan (2006) found no significant difference in classification accuracy between patients with low AHI (<10) and high AHI (>10) index. In our results however, kappa of the WAKE-REM-NREM classification problem was significantly correlated with the AHI index ($r = -0.682$; $p < 0.001$). An increased amount of apneic events during the night thus has a negative impact on the obtainable accuracy of the sleep stage classification problem. This raises the question whether a reliable distribution of sleep stages is even necessary when a significant amount of apneic events occur. Diagnosis of sleep apnea is in that case of most concern, since its negative impact on sleep continuity and depth is anyhow implied.

5. Conclusions and future work

This study evaluated and discussed the use of cardiac and respiratory information for detection of apneic breathing, and for discrimination between sleep stages. Analysis was performed on the UCD sleep apnea database, available on Physionet. While the newly proposed probabilistic representation of the hypnocorrogram and apneacorrogram allows to put classifier output in perspective, further improvements in feature definition (e.g. features related to cardiorespiratory coupling) would be necessary to make a cardiac and respiratory based classifier fully reliable for use on patients with sleep apnea. Detection of apneic breathing proved to be more difficult when compared with the Apnea-ECG database (also available on Physionet), related to a higher proportion of hypopneas and differences in expert apneic annotations. Future work should investigate the impact of automatically including recovery breaths to the expert apneic annotations. Further on, the presence of apneic events proved to have a significant influence on the performance of a cardiac and respiratory based algorithm for sleep stage discrimination. Future work should investigate whether including movement and features related to cardio-respiratory coupling can significantly improve overall accuracy and reliability, and whether epochs which

overlap with apneic events should be ignored during the training phase of the classifier.

Acknowledgments

Research supported by Research Council KUL: CoE PFV/10/002 (OPTEC), PhD/Postdoc grants; FWO: projects G.0427.10N (Integrated EEG-fMRI), G.0108.11 (Compressed Sensing), G.0869.12N (Tumor imaging), G.0A5513N (Deep brain stimulation), PhD/Postdoc grants; ACD is a postdoctoral fellow of the Research Foundation Flanders (FWO); IWT: projects TBM 080658-MRI (EEG-fMRI), TBM 110697-NeoGuard, PhD/Postdoc grants; iMinds Medical Information Technologies: SBO 2015, ICON NXT_Sleep; Flanders Care: Demonstratieproject Tele-Rehab III (2012-2014); Belgian Federal Science Policy Office: IUAP P7/19 (DYSCO, 'Dynamical systems, control and optimization', 2012-2017); Belgian Foreign Affairs-Development Cooperation: VLIR UOS programs; EU: The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP7/2007-2013) / ERC Advanced Grant BIOTENSORS (n 339804). This paper reflects only the authors' views and the Union is not liable for any use that may be made of the contained information. RECAP 209G within INTERREG IVB NWE program, EU MC ITN TRANSACT 2012 (n 316679), ERASMUS EQR: Community service engineer (n 539642-LLP-1-2013).

References

Bsoul, M., Minn, H. & Tamil, L. (2011). Apnea medassist: real-time sleep apnea monitor using single-lead ecg, *Information Technology in Biomedicine, IEEE Transactions on* **15**(3): 416–427.

Burgess, H. J., Trinder, J., Kim, Y. & Luke, D. (1997). Sleep and circadian influences on cardiac autonomic nervous system activity, *American Journal of Physiology-Heart and Circulatory Physiology* **273**(4): H1761–H1768.

Caruana, R. & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms, *Proceedings of the 23rd international conference on Machine learning*, ACM, pp. 161–168.

de Chazal, P., Penzel, T. & Heneghan, C. (2004). Automated detection of obstructive sleep apnoea at different time scales using the electrocardiogram, *Physiological measurement* **25**(4): 967–983.

Himanen, S.-L. & Hasan, J. (2000). Limitations of rechtschaffen and kales, *Sleep medicine reviews* **4**(2): 149–167.

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data, *biometrics* pp. 159–174.

Malik, M., Bigger, J. T., Camm, A. J., Kleiger, R. E., Malliani, A., Moss, A. J. & Schwartz, P. J. (1996). Heart rate variability standards of measurement, physiological interpretation, and clinical use, *European heart journal* **17**(3): 354–381.

Morgenthaler, T., Alessi, C., Friedman, L., Owens, J., Kapur, V., Boehlecke, B., Brown, T., Chesson Jr, A., Coleman, J., Lee-Chiong, T. et al. (2007). Practice parameters for the use of actigraphy in the assessment of sleep and sleep disorders: an update for 2007., *Sleep* **30**(4): 519–529.

Pan, J. & Tompkins, W. J. (1985). A real-time qrs detection algorithm, *Biomedical Engineering, IEEE Transactions on* **32**(3): 230–236.

Penzel, T. & Conradt, R. (2000a). Computer based sleep recording and analysis, *Sleep medicine reviews* **4**(2): 131–148.

Penzel, T., Moody, G., Mark, R., Goldberger, A. & Peter, J. (2000b). The apnea-ecg database, *Computers in Cardiology 2000*, IEEE, pp. 255–258.

Peppard, P. E., Young, T., Barnet, J. H., Palta, M., Hagen, E. W. & Hla, K. M. (2013). Increased prevalence of sleep-disordered breathing in adults, *American journal of epidemiology* **177**(9): 1006–1014.

Pudil, P., Novovičová, J. & Kittler, J. (1994). Floating search methods in feature selection, *Pattern recognition letters* **15**(11): 1119–1125.

Quan, S., Gillin, J. C., Littner, M. & Shepard, J. (1999). Sleep-related breathing disorders in adults: Recommendations for syndrome definition and measurement techniques in clinical research. editorials, *Sleep* **22**(5): 662–689.

Rechtschaffen, A. & Kales, A. (1968). *A manual of standardized terminology, techniques and scoring system for sleep stages of human subjects*, US Government Printing Office, US Public Health Service.

Redmond, S. J. & Heneghan, C. (2006). Cardiorespiratory-based sleep staging in subjects with obstructive sleep apnea, *Biomedical Engineering, IEEE Transactions on* **53**(3): 485–496.

Ripley, B. D. (1996). *Pattern recognition and neural networks*, Cambridge university press.

Schulz, H. (2008). Rethinking sleep analysis, *J Clin Sleep Med* **4**(2): 99–103.

Somers, V. K., Dyken, M. E., Clary, M. P. & Abboud, F. M. (1995). Sympathetic neural mechanisms in obstructive sleep apnea., *Journal of Clinical Investigation* **96**(4): 1897–1904.

Somers, V. K., Dyken, M. E., Mark, A. L. & Abboud, F. M. (1993). Sympathetic-nerve activity during sleep in normal subjects, *New England Journal of Medicine* **328**(5): 303–307.

St. Vincent's University Hospital/University College Dublin (2008). Sleep apnea database. `http://www.physionet.org/pn3/ucddb/` [Last accessed: 2014-11-07].

Verhaert, V., Haex, B., De Wilde, T., Berckmans, D., Vandekerckhove, M., Verbraecken, J. & Sloten, J. V. (2011). Unobtrusive assessment of motor patterns during sleep based on mattress indentation measurements, *Information Technology in Biomedicine, IEEE Transactions on* **15**(5): 787–794.

Whitney, C. W., Gottlieb, D. J., Redline, S., Norman, R. G., Dodge, R. R., Shahar, E., Surovec, S. & Nieto, F. J. (1998). Reliability of scoring respiratory disturbance indices and sleep staging., *Sleep* **21**(7): 749–757.

Willemen, T., Van Deun, D., Verhaert, V., Vandekerckhove, M., Exadaktylos, V., Verbraecken, J., Huffel, S., Haex, B. & Vander Sloten, J. (2014a). An evaluation of cardiorespiratory and movement features with respect to sleep-stage classification, *IEEE journal of biomedical and health informatics* **18**(2): 661–669.

Willemen, T., Varon, C., Haex, B., Vander Sloten, J. & Van Huffel, S. (2014b). Assessment of different methodologies to include temporal information in classifying episodes of sleep apnea based on single-lead electrocardiogram, *Proc. of the 41st Computers in Cardiology Conference*, pp. 1–4.

Xie, B. & Minn, H. (2012). Real-time sleep apnea detection by classifier combination, *Information Technology in Biomedicine, IEEE Transactions on* **16**(3): 469–477.